Date: 7/30/01 Express Mail Label No. EL552284709US

Inventor(s): Jonathan Stern, Jeremy W. Rothman-Shore, Kosmas Karadimitriou and Michel Decary

Attorney's Docket No.: 2937.1000-008

DATA MINING SYSTEM

RELATED APPLICATIONS

This application claims the benefit of U.S. Provisional Application No. 60/221,750, filed on July 31, 2000, the entire teachings of which are incorporated herein by reference. This application also relates to U.S. Patent Application No. 09/704,080, filed November 1, 2000; U.S. Patent Application No. 09/703,907, filed November 1, 2000; U.S. Patent Application No. 09/768,869 filed January 24, 2001; U.S. Patent Application No. 09/821,908 filed March 30, 2001; and U.S. Patent Application No. _____, filed July 20, 2001, entitled "Computer Method and Apparatus for Extracting Data from Web Pages", Attorney Docket No. 2937.1000-005, all by the Assignee of the present invention and herein incorporated by reference.

BACKGROUND OF THE INVENTION

Generally speaking a global computer network, e.g., the Internet, is formed of a plurality of computers coupled to a communication line for communicating with each other. Each computer is referred to as a network node. Some nodes serve as information bearing sites while other nodes provide connectivity between end users and the information bearing sites.

The explosive growth of the Internet makes it an essential component of every business, organization and institution strategy, and leads to massive amounts of information being placed in the public domain for people to read and explore. The type

of information available ranges from information about companies and their products, services, activities, people and partners, to information about conferences, seminars, and exhibitions, to news sites, to information about universities, schools, colleges, museums and hospitals, to information about government organizations, their purpose, activities and people. The Internet became the venue of choice for every organization for providing pertinent, detailed and timely information about themselves, their cause, services and activities.

The Internet essentially is nothing more than the network infrastructure that connects geographically dispersed computer systems. Every such computer system may contain publicly available (shareable) data that are available to users connected to this network. However, until the early 1990's there was no uniform way or standard conventions for accessing this data. The users had to use a variety of techniques to connect to remote computers (e.g. telnet, ftp, etc) using passwords that were usually site-specific, and they had to know the exact directory and file name that contained the information they were looking for.

The World Wide Web (WWW or simply Web) was created in an effort to simplify and facilitate access to publicly available information from computer systems connected to the Internet. A set of conventions and standards were developed that enabled users to access every Web site (computer system connected to the Web) in the same uniform way, without the need to use special passwords or techniques. In addition, Web browsers became available that let users navigate easily through Web sites by simply clicking hyperlinks (words or sentences connected to some Web resource).

Today the Web contains more than one billion pages that are interconnected with each other and reside in computers all over the world (thus the term "World Wide Web"). The sheer size and explosive growth of the Web has created the need for tools and methods that can automatically search, index, access, extract and recombine information and knowledge that is publicly available from Web resources.

The following definitions are used herein.

Web Domain

Web domain is an Internet address that provides connection to a Web server (a computer system connected to the Internet that allows remote access to some of its contents).

URL

URL stands for Uniform Resource Locator. Generally, URLs have three parts: the first part describes the protocol used to access the content pointed to by the URL, the second contains the directory in which the content is located, and the third contains the file that stores the content:

<protocol> : <domain> <directory> <file>

For example:

http://www.corex.com/bios.html

http://www.cardscan.com/index.html

http://fn.cnn.com/archives/may99/pr37.html

ftp://shiva.lin.com/soft/words.zip

Commonly, the <protocol> part may be missing. In that case, modern Web browsers access the URL as if the http:// prefix was used. In addition, the <file> part may be missing. In that case, the convention calls for the file "index.html" to be fetched.

For example, the following are legal variations of the previous example URLs:

www.corex.com/bios.html

www.cardscan.com

fn.cnn.com/archives/may99/pr37.html

ftp://shiva.lin.com/soft/words.zip

Web Page

Web page is the content associated with a URL. In its simplest form, this content is static text, which is stored into a text file indicated by the URL. However, very often the content contains multi-media elements (e.g. images, audio, video, etc) as well as

non-static text or other elements (e.g. news tickers, frames, scripts, streaming graphics, etc). Very often, more than one files form a Web page, however, there is only one file that is associated with the URL and which initiates or guides the Web page generation.

Web Browser

Web browser is a software program that allows users to access the content stored in Web sites. Modern Web browsers can also create content "on the fly", according to instructions received from a Web site. This concept is commonly referred to as "dynamic page generation". In addition, browsers can commonly send information back to the Web site, thus enabling two-way communication of the user and the Web site.

As our society's infrastructure becomes increasingly dependent on computers and information systems, electronic media and computer networks progressively replace traditional means of storing and disseminating information. There are several reasons for this trend, including cost of physical vs. computer storage, relatively easy protection of digital information from natural disasters and wear, almost instantaneous transmission of digital data to multiple recipients, and, perhaps most importantly, unprecedented capabilities for indexing, search and retrieval of digital information with very little human intervention.

Decades of active research in the Computer Science field of Information Retrieval have yield several algorithms and techniques for efficiently searching and retrieving information from structured databases. However, the world's largest information repository, the Web, contains mostly unstructured information, in the form of Web pages, text documents, or multimedia files. There are no standards on the content, format, or style of information published in the Web, except perhaps, the requirement that it should be understandable by human readers. Therefore the power of structured database queries that can readily connect, combine and filter information to present exactly what the user wants is not available in the Web.

Trying to alleviate this situation, search engines that index millions of Web pages based on keywords have been developed. Some of these search engines have a user-friendly front end that accepts natural languages queries. In general, these queries are analyzed to extract the keywords the user is possibly looking for, and then a simple keyword-based search is performed through the engine's indexes. However, this essentially corresponds to querying one field only in a database and it lacks the multi-field queries that are typical on any database system. The result is that Web queries cannot become very specific; therefore they tend to return thousands of results of which only a few may be relevant. Furthermore, the "results" returned are not specific data, similar to what database queries typically return; instead, they are lists of Web pages, which may or may not contain the requested answer.

In order to leverage the information retrieval power and search sophistication of database systems, the information needs to be structured, so that it can be stored in database format. Since the Web contains mostly unstructured information, methods and techniques are needed to extract data and discover patterns in the Web in order to transform the unstructured information into structured data.

The Web is a vast repository of information and data that grows continuously. Information traditionally published in other media (e.g. manuals, brochures, magazines, books, newspapers, etc.) is now increasingly published either exclusively on the Web, or in two versions, one of which is distributed through the Web. In addition, older information and content from traditional media is now routinely transferred into electronic format to be made available in the Web, e.g. old books from libraries, journals from professional associations, etc. As a result, the Web becomes gradually the primary source of information in our society, with other sources (e.g. books, journals, etc) assuming a secondary role.

As the Web becomes the world's largest information repository, many types of public information about people become accessible through the Web. For example, club and association memberships, employment information, even biographical information can be found in organization Web sites, company Web sites, or news Web sites.

Furthermore, many individuals create personal Web sites where they publish themselves all kinds of personal information not available from any other source (e.g. resume, hobbies, interests, "personal news", etc).

In addition, people often use public forums to exchange e-mails, participate in discussions, ask questions, or provide answers. E-mail discussions from these forums are routinely stored in archives that are publicly available through the Web; these archives are great sources of information about people's interests, expertise, hobbies, professional affiliations, etc.

Employment and biographical information is an invaluable asset for employment agencies and hiring managers who constantly search for qualified professionals to fill job openings. Data about people's interests, hobbies and shopping preferences are priceless for market research and target advertisement campaigns. Finally, any current information about people (e.g. current employment, contact information, etc) is of great interest to individuals who want to search for or reestablish contact with old friends, acquaintances or colleagues.

As organizations increase their Web presence through their own Web sites or press releases that are published on-line, most public information about organizations become accessible through the Web. Any type of organization information that a few years ago would only be published in brochures, news articles, trade show presentations, or direct mail to customers and consumers, now is also routinely published to the organization's Web site where it is readily accessible by anyone with an Internet connection and a Web browser. The information that organizations typically publish in their Web sites include the following:

- Organization name
- Organization description
- Products
- Management team
- Contact information
- Organization press releases

- Product reviews, awards, etc
- Organization location(s)

...etc...

SUMMARY OF THE INVENTION

Information about people is fairly prevalent on the Internet and almost every Web site contains some mentions about people. For example: many sites put up by companies (company Web sites) include information about their management team, their Public Relations person and in some cases their entire staff. Hospitals, universities and other academic institution sites tend to list their entire faculty and senior staff along with credentials and areas of specialty. News sites, magazines, newspapers, newsletters and other news and information sources contain articles and news about people. Even if the subject of the article is not about a person the article invariably will contain quotes from people with basic information about the organization they work for and their position or title in the organization. For example, an article about the explosive growth of the Web might contain a quote like: " 'Browser technology is now the foundation of our next generation software.' said William H. Gates, founder and Chairman of Microsoft Corporation in Redmond WA, a leading software company."

All of this data is publicly available in the Web. However, since it is not organized in any standard fashion, it is extremely difficult for someone to find answers to questions such as: "What are the names of all Marketing Directors of high-tech companies in the New England area?" The purpose of the present invention is to extract this kind of public data about people from the Web and organize it into a database, so that simple database queries can answer such questions.

In addition to people information, this invention also extracts from the Web organization information. Many people are working in positions that directly relate to the organization's core activities. Hence their skills, knowledge and specialty likely match the activity of the workplace. Gathering information about the organization adds

another dimension to the biographical information collected and maintained about people.

An organization's Web site contains a lot of information about the organization, its business, products, mission, people, location, partners and more. As with people, the described invention can only collect and organize information that exist on the site itself, hence the level, accuracy and amount of collected information will vary from organization to organization. In general, one can expect to find some or all of the following information.

- The full name of the organization and commonly used aliases of it
- The address of the organization headquarters and other offices and subsidiaries. The location of the organization is of great significance since it generally points to the location of its people and therefore augments the record of the people associated with the organization.
- Contact information including phone, fax and certain general email addresses such as sales@corpx.com
- Organization description
- Organization mission
- Products and services
- Common noun phrases. Noun phrases that appear often on an organization Web site are significant sources of information identifying the main keywords describing the organization and hence the people who are associated with it. Noun phrases such as "signal processing", "public relations", "intellectual property" or "early childhood" can dramatically narrow a search for people in a specific profession.

Creating a database about people and organizations could, of course, be done manually. Since this data is publicly available, human employees could scan the Web and other sources and populate a database with the data. However, there are significant drawbacks in this manual approach:

a) it is too expensive. Tens or hundreds of workers would be needed to scan and extract data even for a small fraction of the Web.

b) it is too slow. Scanning and extracting data from one Web site may require many man-hours of work - even working on a single Web page may require several minutes of work.

c) it is error prone. Human errors are unavoidable, both in finding the data and in transferring them to the database.

In contrast to the manual approach, the purpose of the present invention is to develop an automated approach for the data extraction and collection. The benefits over the manual approach are obvious:

a) automation is cheap. Computers can work 24 hours a day, 7 days a week. A single personal computer can replace tens of human workers in this data extraction task.

b) computers are fast. In general, using the method described in this invention, 5 minutes of computer time in a low-cost computer can produce the same data as several man-hours of manual work.

c) very high accuracy can be achieved. Even though programming errors are unavoidable, these errors can usually be found and corrected fairly easily, so that the accuracy of the system increases over time.

In a preferred embodiment, a computer automated system and method mines, from a global computer network, information on people and organizations. The invention system (and method functions/operations) includes:

a plurality of automated crawlers for transversing sites of a global computer network and retrieving pages that contain information of interest;

a distributor coupled to the crawlers for controlling crawler processing;

an extractor responsive to the crawler retrieved pages and extracting information about people and organizations therefrom, the extracted information being stored in a database;

an integrator coupled to the database for resolving duplicate information and combining related information in the database; and

a post processor coupled to the database for analyzing contents of the database and generating missing information therefrom.

Preferably the database stores information about different people in different respective records. Given two records of potentially the same person, the integrator combines the two records if: (a) name of the person is the same in the two records, and (b) either affiliated organization name or respective title is the same in the two records. The integrator may also consider person name - title combination matches in light of the statistical rarity of the title and person's name.

In accordance with one aspect of the present invention, the post-processor generates an email address (e.g. business/non-personal email address) of a subject person named in the database with respect to organization named in the database for the subject person. The email address is generated by the post-processor:

obtaining a working e-mail address to the respective organization, the working e-mail address not being the e-mail address of the subject person;

deducing from the working e-mail address, format of e-mail addresses to the respective organization;

using the deduced information, constructing potential (i.e. candidate) e-mail addresses for the subject person; and

verifying each constructed potential e-mail address by testing each, such that at least one verified constructed potential e-mail address provides a business e-mail address of the subject person.

The post-processor also may utilize predefined common email address formats to construct potential business/non-personal email addresses of the subject person.

BRIEF DESCRIPTION OF THE DRAWINGS

The foregoing and other objects, features and advantages of the invention will be apparent from the following more particular description of preferred embodiments of

the invention, as illustrated in the accompanying drawings in which like reference characters refer to the same parts throughout the different views. The drawings are not necessarily to scale, emphasis instead being placed upon illustrating the principles of the invention.

Fig. 1 is a block diagram of a preferred embodiment of the present invention.

Fig. 2 is a schematic illustration of a global computer network in which the invention system of Fig. 1 operates.

Fig. 3 is a flow diagram of email address interpolation by a post-processor of the embodiment of Fig. 1.

Fig. 4 is a flow diagram of duplicate record detection for merger by an integrator of the embodiment of Fig. 1.

DETAILED DESCRIPTION OF THE INVENTION

As illustrated in Fig. 1, the main components of the invention system 40 are the following:

a) The "Crawler" 11: a software robot that visits and traverses Web sites in search of Web pages that contain information of interest

b) The "Distributor" 47: a software system that controls several Crawler processes

c) The "Extractor" 41: a software module that processes Web pages returned by the Crawler 11 to extract the information about people and companies (organizations).

d) The "Loader" 43: a software program that loads the data found by the Extractor into the database.

e) The database 45: the place where all the information are stored.

f) The "Integrator" 49: a software module that resolves duplicates, and combines related information in the database.

g) The "Post-Processor 51" which enhances the data, in particular analyzes the data and adds missing pieces of information, such as email addresses.

It is understood that each component 11, 47, 41, 43, 45, 49, 51 is implemented in hardware, software or a combination thereof and is executed by digital processing means (e.g., a computer) 27. A single computer or a series of computers processing in parallel, distributed or other fashion is suitable. For example as illustrated in Fig. 2, computer 27 executes invention system 40 in working memory. Computer 27 is coupled across communication lines 23 to a global network 21 of computers 25. Each node 25, 27 on the network 21 has a respective architecture (e.g., local area network, wide area network, client server, etc.) which may use routers, high speed connections, and the like to couple to global network 21. Some nodes may serve as service providers or host servers to a multiplicity of end users, and so forth.

Returning to Fig. 1, the Crawler 11 is a software robot that systematically visits and traverses Web sites in order to identify and collect Web pages that contain information of interest to the users. Such a robot for extracting information about people and organizations is described in detail in U.S. Patent Application 09/821,908 filed March 30, 2001 entitled "Computer Method and Apparatus for Collecting People and Organization Information from Web Sites" by Jonathan Stern, Kosmas Karadimitriou, Jeremy W. Rothman-Shore and Michel Decary.

In order for the Crawler 11 to be effective in collecting Web pages with useful content, it must be able to perform the following functions:

a) identify the type of the site visited (e.g. company Web site, University Web site, personal Web site, etc.)

b) identify the name of the copyright owner of the site (i.e. the individual or organization that is responsible for the content published in the Web site)

c) identify the type of content that individual Web pages contain (e.g. contact information, list of people, multimedia content, etc.)

d) identify early in the process of traversing the Web site its expected structure (e.g. organization and type of Web pages and site link structure)

e) "prune" the site tree in order to avoid visiting parts that are not expected to produce much useful information.

Each of these is addressed in the related applications.

All of these functions are essential so that the Crawler 11 can harvest the most useful and content-rich Web pages from a Web site by visiting as few pages as possible. In other words, the Crawler 11 traverses the site in such a way so that it visits many pages with high information content, and few pages with little or no useful information. Note that the term "useful information" is relevant to what is considered useful by the system users. So, for example, users that use the system to collect information about a chemical process may consider "useful" any Web page with content that relates directly or indirectly to chemistry. On the other hand, users that want to collect information about company locations may consider useful any page that contains at least one address.

All of the data collected by the crawler 11 (web site type, copyright owner, list of interesting pages, etc.) are passed to the other components of the system so that they may use this data in their own analyses.

In general, the automated system 40 described by this invention needs to be as efficient as possible because of the sheer size of Web. One of the measures of efficiency is the number of Web sites visited and traversed per hour. The Web currently contains many million Web sites (estimates in January 2000 set this number to over 10 million Web sites and they keep increasing exponentially). A system that can visit and extract information from an average of 10 Web sites per hour will need at least 1 million hours to cover the entire Web, that is, about 100 years! On the other hand, a system that can visit 1,000 Web sites per hour (100 times more efficient) will need about 1 year to cover the entire Web, whereas a system capable of visiting 10,000 Web sites per hour can cover the entire Web in less than 2 months. These estimates are tabulated in the next table:

System throughput	Time Needed to Cover the Entire Web
10 Web sites/hour	100 years

100 Web sites/hour	10 years
1,000 Web sites/hour	1 year
10,000 Web sites/hour	less than 2 months

The term "system throughput" is used to refer to the number of Web sites visited and processed per hour. The system throughput is related to the average time that the invention system 40 needs to visit and process (extract information) from one Web site:

$$System throughput = \frac{1}{Average\ time\ per\ Website}$$

In general, the average time per Web site is the sum of the times needed by each system module to perform its functions, that is:

$$\text{Average time per Web site} = T_{Crawler} + T_{Extractor} + T_{Loader} + T_{Integrator}$$

where

$T_{Crawler}$ is the time required by the Crawler 11 to crawl one Web site,

$T_{Extractor}$ is the time required by the Extractor 41 to extract data from the contents of one Web site,

T_{Loader} is the time required by the Loader 43 to load in the database 45 the data extracted by the Extractor 41 from one Web site, and

$T_{Integrator}$ is the time required by the Integrator 49 to post-process and clean up the data related to one Web site.

However, by arranging these modules in a pipeline fashion we can process multiple Web sites simultaneously, that is, while the Crawler 11 is crawling one Web site, the Loader 43 is loading the data from another Web site, and the Integrator 49 is processing the data from yet another Web site. In that case, the average time per Web site becomes equal to the maximum time among all components, that is:

Average time per Web site = Maximum of ($T_{Crawler}$, $T_{Extractor}$, T_{Loader}, $T_{Integrator}$)

In any modern computer system, processing data and performing database transactions is much faster than accessing the Web. Even when high-speed Internet connections are used, the time to access and download Web pages is bounded by the speed of the Web servers where the pages reside, by the network "path" that is used to transfer the data, and by the Internet "status" when this communication take place. So in general, accessing the Web is by far the slowest operation that the invention system 40 must perform, therefore the average time per Web site is practically equal to the time required by the Crawler 11 to visit and traverse one Web site:

Average time per Web site = $T_{Crawler}$

and the system throughput becomes:

$$System\ throughput = \frac{1}{T_{Crawler}}$$

This means that in order to increase the system 40 efficiency and be able to traverse the entire Web in a reasonable amount of time, one needs to decrease the average time required to crawl one Web site. However, as noted above, this cannot be achieved by simply using more powerful computers or better Internet connections, because this time is bounded by external factors (responsiveness of external Web servers, Internet status, etc). In other words, no matter how efficiently one builds the Crawler 11 or how fast an Internet connection one uses, the average time that is needed to crawl one Web site cannot be reduced arbitrarily. Thus, the only way to achieve the desired system throughput is to use multiple Crawlers 11. In fact, even when using a relatively slow crawler that visits only 100 sites per hour and it would take by itself 10 years to cover the entire Web, when employing 100 such Crawlers 11 one can cover the entire Web in less than 2 months.

This discussion has demonstrated so far that the only way to achieve reasonable system throughput is to use several (in the order of hundreds) Crawler 11 processes simultaneously. In this case, controlling and administrating manually all these processes becomes a challenge, to say the least. Therefore a separate software module is needed that automates this process. This module is the Distributor 47.

Because the Distributor 47 is integrated with the Crawler 11, it is able to adjust the schedule of which websites to visit by leveraging the information that the Crawler 11 extracted during previous visits. Because the crawler 11 is automatically determining the site type, the Distributor 47 is able to give higher priority to some sites and lower priority to others. For example, if the crawler 11 found a site with a daily news feed, the Distributor 47 may adjust the schedule to visit this website on a daily basis. At the same time, if the crawler 11 finds a website that is uninteresting because it does not contain data relevant to what is being extracted, such as someone's personal website, the Distributor 47 adjusts the schedule to visit the website only every six months or longer.

To summarize, several Crawler 11 processes are needed by the system 40 in order to increase its efficiency, and an automated method must be employed to manage all these processes. The Distributor 47 offers exactly this functionality: it is a software module whose main function is to control and distribute work to multiple Crawlers 11. The Distributor 47 uses a database 14 to keep track of all the Web sites that must be visited, and the visiting schedule for each one (some Web sites must be visited more frequently than others, depending on how often their contents change). In addition, the Distributor 47 prioritizes the Web sites according to their relative importance for the users, and it manages the Crawlers 11 so that the most important sites are visited first. The Distributor 40 is responsible to start multiple Crawler 11 processes, and keep their number as high as possible, without hurting the overall system performance. It also monitors the status of the running Crawler processes and stops or kills any processes that exhibit unwanted behavior (e.g. a process that takes too long, uses too much memory or disk space, etc).

Every Crawler 11 returns and saves in local storage 48 a set of Web pages 12 that potentially contain useful information. These Web pages 12 are then processed by a software module that can extract data from HTML code or plain text, the Extractor 41.

The Extractor 41 uses linguistic methods to parse and "understand" text so that it identifies and extracts useful information. The users of the system 40 define what they consider to be "useful" information, and customize accordingly the Extractor 41. Note that the Extractor 41 itself is a very generic and flexible tool, that has the ability to read and parse correctly text written in any language, according to the syntax and grammar rules of that language. However, it needs customization in order to work with a specific language (e.g. English, French, German, etc.) and furthermore, it needs to be "trained" in order to recognize what the users consider useful information.

Customizing the Extractor 41 for a specific language means that one provides it with a set of syntax and grammar rules so that it correctly identifies subject, verb and object in a sentence, it recognizes the time that the sentence refers to (past, present or future), it recognizes the beginning and end of sentences, etc. Training the Extractor 41 for recognizing "useful" information means that one provides it with rules and dictionaries of specific terms so that it recognizes keywords and using the rules it decides when something is useful or not. In general, this training may be automated in a significant level, by using examples of "useful" and "useless" text and let the Extractor 41determine statistically what are the terms that may be considered as keywords of useful information, and also what are good rules (or tests) that may be used during the data extraction process. There are various methods and techniques that the Extractor 41 may use for its internal decision making and pattern recognition, for example, template-based pattern recognition (see U.S. Patent Application No. 09/585,320 filed on June 2, 2000 for a "Method and Apparatus for Deriving Information from Written Text"), Bayesian Networks for decision making (see U.S. Patent Application No. 09/704,080, filed November 1, 2000 entitled "Computer Method and Apparatus for Determining Content Owner of a Web Site"; U.S. Patent Application No. 09/703,907, filed November 1, 2000 entitled "Computer Method and Apparatus for Determining Site

Type of a Web Site; U.S. Patent Application No. 09/768,869 filed January 24, 2001 entitled "Computer Method and Apparatus for Determining Content Types of Web Pages"), Neural Networks for data classification, Genetic Algorithms for selection of "good" rules and keywords, etc.

Where the Extractor 41 is part of the same system 40 as the Crawler 11, it uses the list of interesting pages 12 (stored at 48 in Fig. 1) as input to process on. Also, it uses the website type and determined copyright owner to assist interpretation of the data. For example, if the website is a Company website, Extractor 41 concludes people whose names are found on a management team page on the website work for the company identified as the copyright owner, even though such is not directly stated on the page.

For a detailed description of a preferred Extractor 41 that is customized to extract information about people from the Web see U.S. Patent Application No. _____, filed July 20, 2001 entitled "Computer Method and Apparatus for Extracting Data from Web Pages", Attorney Docket No. 2937.1000-005. That Extractor 41 uses various methods and techniques described in U.S. Patent Application No. 09/585,320 filed on June 2, 2000 for a "Method and Apparatus for Deriving Information from Written Text".

For mining information about people and organizations, the Extractor 41 extracts the following data:

a) Names of People

b) Positions that these people hold or have held, including title, organization name, organization location, state and end dates, and whether the person still holds the position

c) Educational degrees these people have received

d) Certifications that these people have received (e.g. CPA, RN, LCSW, ...)

e) An email address for the person, if available

f) A description of the copyright owner, if it is a company website

g) An address for the copyright owner, if it is a company website

h) Subsidiaries, partners and competitors for the copyright owner, if it is a company website

i) Number of employees

j) Relevant date for each piece of information - some documents are old, even if they are recently published. If a document is dated, the date must be collected and attached to all information found on it.

In the preferred embodiment, Extractor 41 places the foregoing extracted data into working records 16 (for information on people), 17 (for information on organizations).

After the Extractor 41 has processed the Web pages returned by a Crawler 11 and it has extracted the useful information, it passes the extracted information (records 16, 17) to the Loader 43, which is the software module responsible for storing the information in the database 45. One of Loader's 43 responsibilities is to make sure that the information is internally consistent, for example, with no duplicate or conflicting data (i.e., no duplicate records 16, 17). The Loader 43 also implements data filtering rules that have been given by the system users in order to avoid cluttering the database with "garbage" data. For example, in a system built to collect people information, the Extractor 41 may return any information it finds connected to a person's name. However, the Loader 43 may employ filters to discard any information referring to fictional characters or historical figures, e.g. Donald Duck or Alexander the Great, and load in the database 45 only what appears to be current information about real (and alive) people.

Note that some of this filtering may also be performed as post-processing by the Integrator 49, however, by doing the filtering prior to loading the information/records 16, 17 into the database 45, one avoids cluttering the system with obviously useless data. Also some filtering rules may be employed by the Extractor 41 , however, the Extractor 41 preferably does not communicate directly with the database 45, and some of the filtering may require database access.

Another major responsibility of the Loader 43 is to merge information found by the Extractor 41 from separate Web pages in the same Web site. In general, the Extractor 41 works in a page-by-page mode, extracting any information it finds in each individual page. Very often though, the same information may be found repeatedly in more than one Web page from the Web site, e.g. every press release potentially contains the company address. The Extractor 41 itself may keep track of what information it has found as it progressively process all the Web pages from a Web site, however, that would require a lot of "bookkeeping" from the Extractor 41. A simpler way is to let the Extractor 41 extract all the useful information it finds, and then let the Loader 43 decide what is duplicate information or merge pieces of partial information. For example, the first and last name of an employee may be found in one Web page, whereas another Web page contains only his last name and his title. The Loader 43 recognizes that these two pieces of information actually complement one another, and that they may be safely merged into one piece that contains the first name, last name, and title of the person.

The Loader 43 also uses the data collected by the Crawler 11 and the Extractor 41 to tie disparate pieces of information together at the database level. For example, if the Crawler 11 finds that the owner of a website is company "A", and the Extractor 41 finds an address for company "A" and a person working for company "A", the loader 43 combines this information when storing it in the database 45 to show that this person works for company A at the found address.

The Loader 43 also assigns a date to all of the information that it loads. A press release is often maintained on an organization's Web site for years, but the information can quickly go out of date. For example, if the CEO of a company is replaced, all of the older press releases will still refer to that person as working at the company. In any kind of news article or press release, the date of the information must be captured.

Each record 16, 17 carries two dates: the date that the information was extracted, and the original date of the document if such a date can be found. If a page does not contain a date and it is a management team page, it can generally be assumed to be current.

The modification date of a document on the Internet cannot be used to date the information in the article, since this can change for technical reasons, such as using a new layout for a Web site, a page can be dynamically generated, etc.

A preferred embodiment of Loader 43 is described in the related U.S. Patent Application No. _____, filed July 20, 2001, entitled "Computer Method and Apparatus for Extracting Data from Web Pages", Attorney Docket No. 2937.1000-005, cited above.

The database 45 used by the system 40 must be a modern high-end database that can handle large quantities of data and a high number of transactions. The amount of data collected by the system 40 can potentially tax the capacity and capabilities of any database system, therefore particular attention must be paid to the specifications and maintenance of the database 45. Of course, this also depends on the user requirements and the type of information that the system is designed to collect; for example, a system that collects information about "the health industry" probably requires higher capacity database than another system that collects information about "zebras". In addition, a system that offers a Web interface through which anybody in the world may access the data requires a much more powerful database than another system which is not expected to have more than 10 users at a time browsing through the data.

Another part of the invention 40 system is the Integrator 49, the software module that periodically operates on the data in the database 45 trying to identify duplicate data, aliases, and merge or remove any incomplete or low-quality data. In essence, the Integrator 49 finds and exploits any data connections that may exist in the database 45.

When performing Web data mining, there are basically three types of "data connections" or associations:

a) Data connections within a Web page

An example of this type of data connection is a Web page that contains the biography of a person. This page may start with the sentence "When Mr. Jonathan Stern,

CEO of Corex Technologies Corp., decided to..." From this sentence, the following data can be extracted:

"Mr. Jonathan Stern, CEO, Corex Technologies Corp., present"

Later on the same page, another sentence may contain: "Prior to Corex, Jonathan was the CEO of Rosh Intelligent Systems where he..." From this sentence, the following data can be extracted:

"Jonathan, CEO, Rosh Intelligent Systems, past"

It is obvious to a human reader that these two pieces of data are interconnected, referring to the same person. This is a very common type of data interconnections in text that was meant to be read by humans. It assumes that as the reader proceeds through the text he/she keeps a mental "trace" or "memory" of the information already given so that there is no need to repeat continuously in every sentence "Mr. Jonathan Stern, CEO, Corex Technologies Corp".

In the system 40 described in the current invention, this level of data interconnections are handled and resolved at the Extractor 41 level.

b) Data connections within a Web site

Another type of data connection may exist in the Web site level. Very often, a Web site is focused on providing information about a specific "subject". For example, company Web sites usually provide information related to the company, whereas a Web site maintained by the "Johnny Cash Fan Club" probably contains information that is focused exclusively to the singer Johnny Cash. A Web site with such a strong focus tends to assume that human readers are familiar with the central subject of the site and so the site often provides incomplete information in its Web pages. For example, a company Web site may provide the company address in some Web page without including the company name, since it is assumed that a human reader already knows the company name.

As it has been described in the previous sections, these type of data connections are handled in the system by the Loader 43.

c) Data connections among different Web sites

Finally, the third type of data connections refers to data collected from different Web sites. For example, in a system built to collect company information, the products of "RND Corporation" may be found in the RND Corporation's Web site, the stock ticker for this company may be found in the Fidelity Investments Web site, a brief description about the company may be found in a press release from the PRNewsWire Web site, whereas reviews about the company's flagship product may be found in a trade publication's Web site.

In the current system 40, the Integrator module 49 identifies and handles this type of data connections. As the database 45 is populated with new data and older data are "refreshed" by revisiting Web sites, new interconnections of this type are continuously introduced. For every new piece of information, the Integrator 49 "scans" the database to find other pieces of information that potentially share a connection. Fig. 4 illustrates the Integrator 49 process.

Combining people information from different sources is performed only if a reasonable match is found on two separate pieces of data between two records 16. One of the pieces of data needs to be the name of the person (step 121), since two people with different names are almost never the same people. However, name matching alone is not good enough, since many different people in world share the same name. Thus, supporting evidence needs to be found.

At step 123, if the two records 16 of people with the same name are found working for the same company (organization), either currently or in the past, the records 16 are assumed to be of the same person and therefore combined (at 140). This is almost always safe for smaller companies. For very large companies, it is possible that there are two people with the same name working there. However, the chances that both of these people are mentioned on the web and found by the system 40 reduce the

chances that this situation will actually be encountered. Thus, while this process potentially may introduce a small amount of erroneous combinations, the vast majority will be correct.

If a match on company names is not found at step 123, titles are used (at step 125) to determine whether subject records 16 of people information should be combined. In this case, if two records 16 of people with the same name are found to have the same title, it is possible to combine the subject records 16. This cannot be done blindly, since it is entirely possible that there are two people named "John Smith" with the title of "Product Manager" in the world. So after step 125 detecting same title, step 127 determines statisical rarity of the title indication shared by the two subject records 16. The database 45 itself may be used to determine statistics of the frequency of titles. Titles that appear very common, such as "Product Manager", would not be combined on, but a relatively rare title, such as "Patent Clerk", would be combined on (at 140), since the chances of two people with the same name and that particular title are very low. Thus, while this will also generate some erroneous combinations, the vast majority will still be correct.

Statistics on names may also be used when combining on a name-title match (step 129). For example, if a relatively rare name, such as "Geoffrey Westerchest" was encountered for two separate records 16, the chances that they are the same person are higher, because there are fewer people out there with that name. Thus, how rare a title needs to be might be relaxed in that case. In other words, while it is quite possible that two different instances of "John Smith, Product Manager" are two different people, it is unlikely that two instances of "Geoffrey Westerchest, Product Manager" are different people. Thus, at step 140, Integrator 49 combines records 16 corresponding to these two Geoffreys, i.e., statistically rare name, but not so rare/uncommon title determined at step 129.

Now that the data is completely integrated, connected pieces of information can be used to interpolate missing data by post-processor 51 (Fig 1). An example of this is

using the email addresses of one or two people at an organization to compute email addresses for the other people in the organization.

Whenever a person is associated with an organization, post-processor 51 attempts to identify an email address for that person. Most sites will list an email address for at least one of the people in the organization, or at the very least a generic email for the site (i.e. sales@corex.com) revealing the domain name used for sending emails, which in some cases might be different than the domain name of the site. As soon as a single email address is found, post-processor 51 deduces email addresses for the rest of the people at the organization as follows and illustrated in Fig. 3.

Most organizations have a standard format for their email addresses based on name of a person. From one found/extracted email address of a person at a given organization, post-processor 51 reverse engineers the organization's standard format for email addresses at step 101. At step 103, the preferred algorithm searches for substrings of the known/given person's name within the given email address. For example, if the name is Dexter Sealy, and his given email address is desealy@corex.com, the last name is completely contained within the email address, and the given email address starts with the first two letters of the corresponding person's (Dexter's) first name. So two patterns are identified (step 105): {first name: 1st 2 characters}{last name}@corex.com and {first name: 1st 2 characters}{last name: 1st 5 characters}@corex.com. From the identified patterns, step 107 forms and applies rules to database records 16 that indicate people at the given organization whose email addresses are missing from the records 16.

Accordingly, the people at the given organization whose records 16 do not indicate respective email addresses have these rules applied to their names, as indicated in the name fields of records 16. As a result, in the foregoing example, from the name Jeremy Rothman-Shore in a record 16 name field, the candidate or potential email addresses jerothm@corex.com, and jerothman-shore@corex.com are produced (output) at step 107 of post-processor 51.

In the event that there is no person's email address from which to reverse engineer the company standard email address format and interpolate (generate) email

addresses for others at the given company/organization (at 113, Fig. 3), or the reverse engineering process fails (some people have emails that do not follow the organization standard email address format) after step 107, the post-processing routine 51 applies preferred rules 111 of the most common combinations for creating an email address. Such common combinations include:

{first}.{last}@{server name}

{last}@{server name}

{first x letters of last name}@{server name}

{1st letter of firstname}{lastname}@{server name}

{lastname}{1st letter of firstname}@{server name}

...etc...

In addition, most email servers will try to alias email addresses to someone in the organization and forward on the message, even if the message originally used an incorrect email address. For example, many email servers will accept an email address in the form of {first name}_{last name}@{server name} and send it to the appropriate person.

In order to verify the interpolated/generated email addresses, step 115 sends a test or email message using a candidate generated email address, out to the person from post processor 51/invention system 40. If the first tested candidate email address is incorrect, an unrecognized recipient reply will be sent back from the mail server to system 40 (host server 27). In such a case, post-processor 51 tries another candidate or an alternate variant of the email address at test step 115 until either a mail delivery acknowledgment is received or no error reply comes back. A unique code may be embedded in the subject field of each trial to simplify matching it with the delivery acknowledgment or error message.

While this invention has been particularly shown and described with references to preferred embodiments thereof, it will be understood by those skilled in the art that

various changes in form and details may be made therein without departing from the scope of the invention encompassed by the appended claims.

For example, reference is made to the Internet and Web sites thereon. The present invention may be applied to other global computer networks and is not dependant on the web platform or HTTP protocol, and the like.

The terms "company" and "organization" are used to refer to a variety of entities and/or employers such as businesses, associations, societies, governmental bodies, clubs and the like. Hence association with anyone of these entities is generically termed "employment" or business/non-personal relations to and is intended to cover any affiliation, membership, or connection a person has with the corresponding entity. That is, the terms "employer" and "employed by" are to be given a more generic interpretation liken to non-personal/business affairs of a person. Similarly the term "business email address" is intended to distinguish from a personal, private, at-home email address of a person, but may correspond to any of the variety of entities noted above.